SUBJECT LINES AS SENSORS: CO-WORD ANALYSIS OF EMAIL TO SUPPORT THE MANAGEMENT OF COLLABORATIVE ENGINEERING WORK

Jones, Simon L. (1); Payne, Stephen J. (1); Hicks, Ben J. (2); Gopsill, James A. (2); Snider, Chris (2); Shi, Lei (1)
1: University of Bath, United Kingdom; 2: University of Bristol, United Kingdom

Abstract

This paper presents a topic-based analysis of email subject line data from a large-scale engineering project and explores its utility for supporting the management of collaborative work. The main contributions of the paper are a novel interpretation of the co-word network analysis method for application within an engineering project management context, and the appraisal of the method for finding patterns within subject line data. Our findings suggest that the approach has the potential to contribute to monitoring work complexity, tracking progress, recognizing synergy and divergence, detecting scope creep, and supporting knowledge capture.

Keywords: Communication, Information management, Project management, Visualisation

Contact:
Dr. Simon Lloyd Jones
University of Bath
Department of Computer Science
United Kingdom
s.l.jones@bath.ac.uk

1 INTRODUCTION

In many organizations, project teams are required to collaborate across departmental, corporate, and geographical boundaries on highly complex, dynamic, and interdisciplinary projects. These projects (e.g. software development, construction and heavy engineering) typically require major planning, coordination, control and monitoring efforts to ensure their success. Early identification of potential problems or deviation from an intended path of progression is essential for successful delivery in terms of time, cost, quality and scope. It follows that management of complex collaborative projects requires awareness about the actions of group members and timely information about project performance, such that decisions affecting the development of the project can be made judiciously. It has been highlighted by many authors, (e.g. Grudin, 1988, Liu & Horowitz, 1989), that tools to support the control of large collaborative projects are a necessity, since group members simply have too much information to keep track of. However, a reliance on group members to manually provide the information required for such support is both error prone (due to omissions, lack of availability, bias and latency in provision of information), and costly, because of the additional effort that is required (Grudin, 1988).

In order to provide a more automated approach to support the management of projects, this paper focuses on email as a valuable source of information for feedback about what is happening in a project. It has been noted (e.g. Mackay, 1988, Ducheneaut & Bellotti, 2001) that email is the core means by which work is received, managed, and delegated, and that email provides a mechanism for supporting a variety of problem-solving and management activities on engineering projects (Wasiak et al., 2010, Loftus et al., 2010, Hicks, 2013); hence a considerable amount of up-to-date information about the state of a project is likely to reside in participants' inboxes. We posit that by using automated analysis to uncover details of the topics of email communications, we can reveal information that is useful to project members in a number of ways, for example for understanding the state of the project, assessing its trajectory and accordance with a schedule, and increasing awareness of emerging issues.

In this paper we investigate the utility of email subject line data, extracted from project-related emails in a 47-month industrial collaborative engineering project, as an additional 'sensor' to support the management of collaborative work. We apply co-word network analysis techniques (Callon et al., 1983) to subject lines in order to elicit the latent topics of project-related communications, and present various visualizations of the data in order to explore their utility for revealing patterns and trends. We attempt to triangulate patterns revealed by the topic analysis with ground-truth data about the project from interviews with project members, reviews of project-related documentation, and manual inspection of the email content.

2 RELATED WORK

Within 'text mining' research literature, numerous authors have investigated topic detection from emails, often with the primary motivation of assisting individual users in dealing with email overload and personal information management (e.g. Surendran et al., 2005, Dredze et al., 2006, Cselle et al., 2007), rather than understanding and managing the state of collaborative projects. Holistic approaches to communication analysis (i.e. analysing entire organisations or projects, rather than individuals) have been employed in recent work. For example, Wolf et al. (2009) showed that social-network analysis of all communications between software development teams could provide insight into potential causes of failure within software development projects. Similarly, analysts have used interaction patterns to interpret and uncover aspects that have been responsible for successful collaborations within groups (for instance during design tasks (Korba et al., 2006)). In engineering design literature, Loftus et al. (2008) highlighted the importance of exploring ways to improve the utility of email for supporting the design process. Hicks (2013) showed that identifying different *types* of email could be useful for revealing signatures that align with project phases and, more importantly, problems encountered, indicating the potential value of email content in a project management context. However, this work relied on manual coding of email, which is labour-intensive and impractical for use in a system intended to provide timely feedback about a project. Jones et al. (2013), Gopsill et al. (2013 & 2014), and Shi et al. (2014) demonstrated the utility of automated analysis of digital objects (such as email, CAD files and project documentation) for understanding the engineering design process. However, automated topic-based content analysis of email communications for supporting the management of engineering projects remains unexplored.

While existing techniques for automatically identifying email topics are many and varied, in this paper we focus on co-word analysis (Callon et al., 1991), a topic detection method that is prevalent in the intellectual mapping of scientific fields (e.g. Coulter et al., 1998, Liu et al., 2014), but which is seldom applied to the analysis of email. Our intention is to produce analogous topic mappings within the context of complex collaborative projects, revealing how projects are structured in terms of their topics, how they are evolving over time and what adjustments project members might need to make, in order to support project management.

3 DATA COLLECTION

The e-mail data analysed in this paper was collected from an industrial collaborative engineering project that lasted 47 months, involved four teams distributed over four different countries, and required the design, manufacture and testing of control and power systems for a customer in the marine sector. The contract required the provision of six subsystems, based on specific requirements of end users.

For litigation purposes, it was part of the Project Director's responsibilities to organise and maintain a central email repository. Staff members were requested to copy project-related mail into the repository, which could automatically organise emails by features such as date or sender, and also allowed mail to be manually tagged and sorted into folders according to project. Hence we were able to extract all stored mail relating to a single project. 10,628 emails were collected from the repository for the project (an average of ~11 emails for each working day), involving a combined total of 1,045 senders and recipients. There were 8,523 unique subject lines in total. 62 (approx. 0.5%) of the email subject lines were left blank. 132 email bodies (1.2%) were left blank.

In addition to the email collection, data was collected through post-project interviews with project stakeholders, and through review of project documentation. Five formal interviews were conducted with participants who were chosen because of their diverse roles in the project, namely; Project Director (responsible for the entire project), Project Manager (responsible for the day-to-day running of systems and software aspects of the project), Engineer (supporting the project manager and working on software development), Commercial Manager (resolving contractual disputes and reviewing correspondence histories), and Warranty Support (responsible for communication with end users over issues arising). The semi-structured interviews each lasted one-hour. Participants were asked to describe how their role changed over time, who they communicated with throughout the project, what they communicated about and why they were communicating (as described in Wasiak, 2010).
Project documentation was reviewed by two researchers and used to provide an overview of the project at different stages. Monthly reports and time plans indicating project activities were of particular interest. Project schedules were used to identify key phases in the project. Minutes and reports from stage review meetings provided further detail. Reviewing project documentation and interviewing project members enabled the reported events of a project to be compared with patterns revealed by co-word topic analysis of email subject lines. This triangulation of sources was intended to provide a perspective on whether useful information could be revealed by the co-word topic analysis – moreover, information that might otherwise only be available through less automatable means of investigation.

4 DATA PREPARATION FOR CO-WORD ANALYSIS

The entire email dataset was time-sliced (i.e. organized into various chronological subsets) in order to perform co-word topic analysis on a month-by-month basis, with the consideration that this might reflect a reasonable time interval at which project members would wish to review progress. Monthly reporting mechanisms are common in many collaborative projects (Grønbæk et al., 1993) and thus informed our unit of analysis. Time slicing was performed in two ways; single month only and cumulative. A single month slice labeled as t=5 represents the fifth month of the project only, whereas the cumulative time slice $t \leq 5$ represents all data up to and including the fifth month, and so on. The purpose of performing single month and cumulative slicing was to provide two perspectives on the data, with the former giving a localized view of what was going on in that month, and the latter giving a view of what was taking place in the project as a whole.
Typically, co-word topic analysis relies on the provision of keywords, and rests on the assumption that these keywords constitute an adequate description of the topics. Our work rests on similar assumptions: that emails frequently contain an informative subject line and that email topics do not stray too far from

the subject line. It is generally accepted that the subject line's intended purpose is to provide a summarization of an email's content. To mitigate errors in this assumption we perform filtering on the subject line data to eliminate words that may not relate to an email's topic: Porter stemming (Porter, 2009) was applied in order to avoid various derivations of the same word being treated as unique. Additionally, email related mark-up was removed (e.g. Fwd:, Re:, cc:, etc.); non alphanumeric characters were removed; all characters were converted to lower case to avoid case sensitivity; numeric words were removed (e.g. dates and times); and common stop-words (e.g. the, with, and, if, etc.) were removed using NLTK 3.0 for Python.

5 TOPIC DETECTION USING CO-WORD ANALYSIS

Co-word analysis involves construction of a network graph representing the co-occurrence relationships between topic-related words. Words that appear within a subject line are therefore treated as co-words of all other words within the same subject line, and form vertices in the network graph, with connecting edges between them. Edges are given weights based on the frequency at which two words co-occur within the entire dataset, normalized by the frequency with which they occur independently (Callon et al., 1991).

Co-word networks were constructed for each of the 94 time slices (47 single month + 47 cumulative). We used a second stage of filtering, based on graph metrics, to reduce the set of keywords even further to those that most reliably capture coherent topics and themes. We removed any keywords that exhibited very low degree centrality, very low frequency, and very low weights on all of their edges. Hence, we ignore any words that only occur once and do not co-occur frequently with many other words. With the co-word network graphs we applied the Louvain method for cluster detection (Blondel et al., 2008) as a means to derive topics. The method unveils hierarchies of clusters and allows granularity adjustments to discover sub-clusters. We were able to identify instances of topic clusters reappearing, evolving, splitting, and merging using a measure of similarity $s_{A,B}$ between two clusters, A and B, at different points in time: $s_{A,B} = \frac{|A \cap B|}{\max(|A|,|B|)}$ Where a single topic has strong similarities to multiple topics from a previous month it suggests that these topics have merged. Where multiple topics all have a strong similarity to a single topic from a previous month it suggests that this topic has now split into discrete sub-topics.

6 RESULTS AND DISCUSSION

Table 1 illustrates the clustering output for a single time slice (in this case the first month of the project, t=1). The words from each cluster reveal the key elements of each topic – which can be interpreted in order to produce a description of each topic. Thirteen unique topic clusters were identified for this month. The output provides insight into the main topics of the project at this time. For example, we can see that the work relates to various technical aspects of the project: e.g. propulsion and thruster systems, power and automation systems, the design, build and installation of a control console, as well as other management and process oriented topics: e.g. compliance plans, equipment ordering, contracts, work breakdown and resource codes. We argue that the breadth of topics captured by the analysis increases its usefulness, particularly when compared to manual reporting mechanisms that are often limited by the assumed relative importance of particular topics by the author, which may restrict the topics reported on. Due to space limitations we do not present all of the topic clusters from all of the time slices of the dataset. Rather, we discuss interesting patterns and trends (termed signatures (Snider et al., 2014)) that correspond with known information about the project (taken from the project documentation and interviews with project members) and explain how they could support awareness of important issues and inform project management.

Table 1. Topic clusters from the co-word network (t=1)

#	Topic Words	Topic Description
0	control, box, consol, network, specif, design, outstat, install, build	Design, specification, build and installation of an outstation control console.
1	propuls, relay, thruster, overall, quality, protect, plan, software, convert	Propulsion and Thruster Systems (including Protection Relays, Converters and Software Quality Plans)

2	work, breakdown, code, resource	Work breakdown and resource codes
3	miscellan, hmi, pms, schedul, signal, fat, i/o, common, verif, vms, sat	Miscellaneous
4	draw, regist, custom	Customer Drawing Register
5	matrix, respons	Responsibility Matrix – clarifying individual's responsibilities
6	vessel, configure, engine, manag	Vessel management and engine configuration
7	manufactur	Manufacturing
8	function, index, power, system, autom, check, summari	Power and Automation Systems
9	proforma, review, sheet, contract, standard	Contract Standard Specification Proforma
10	map, document	Document Map
11	regulatory, statutory, complianc	Statutory and Regulatory Compliance Plan
12	equip, order	Equipment ordering

6.1 Temporal graphs of topic cluster totals

Figure 1a reveals the number of discrete topics that the project members are discussing within each single month (according to our co-word analysis), whereas Figure 1b reveals the accumulated number of discrete topics in the project over time. The various phases of the engineering project are delineated based on evidence from project documentation and interviews, namely; specification, manufacture, sub-system testing, assembly, and final testing.

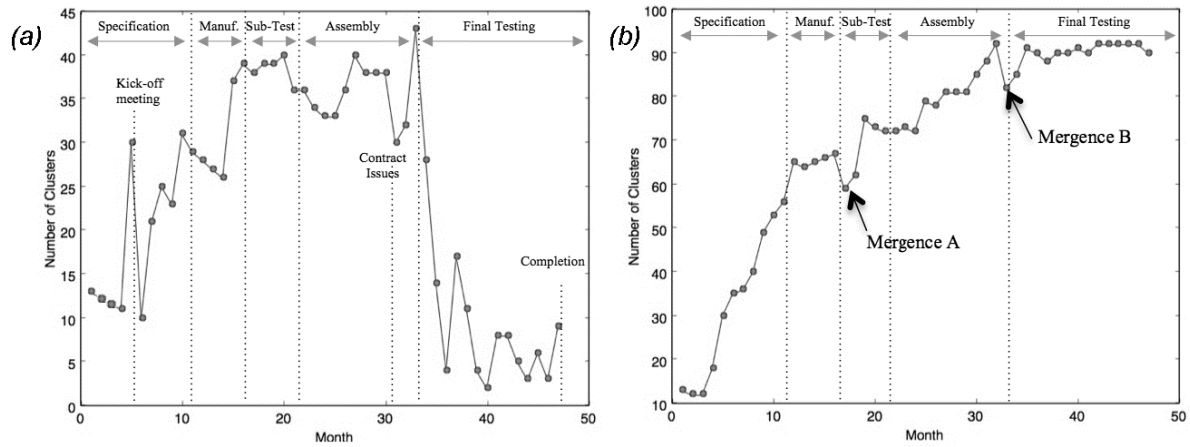

Figure 1. No. of topic clusters each month - (a) single month network slices (b) cumulative co-word network

Signatures of Project Phase Transitions

Both individual month and cumulative measures may be useful as proxies for other features of the project, such as complexity, breadth of focus, focus change, and phase transitions. Changes within these graphs (Fig. 1a & b) closely correspond to documented changes in project phase. Figure 1a shows that the number of topics from one month to the next typically decreases when a phase transition occurs. For example, there are fewer topics at the start of the manufacturing phase than at the end of the specification phase, and likewise for the transitions from manufacturing to sub-system testing, and assembly to final testing. One possible explanation for this trend, based on feedback from interviewees about activities throughout the project, is that a decrease in topics is a manifestation of consolidation work taking place upon completion of a phase, followed by a natural winding-up effect that takes place as a new type of work begins, particularly if there is a handover of responsibility between certain members of the group. There are also some notable step changes within Figure 1a, indicating that certain phases of the project involve a broader or narrower range of topics than others. For example, there are far more topics detected from the late stages of manufacturing through to the completion of assembly, and a sudden drop-off in the number of topics within the final testing phase. This drop off might be expected as during the final testing stage topics should reflect essential requirements of the product specification, and in particular,

a constrained set of topics related to the acceptance and approval of the system. Indicators of phase transitions in subject line data could therefore be useful for verifying if work is on/behind schedule.

Signatures of Divergence and Shifts in Focus

Step changes are likely to be significant indicators of a shift in focus within the project, since they represent dramatic alterations in the topic structure. Sudden peaks or troughs within the graph also correspond with significant events reported in project documentation and interviews. For example, the peak at month 5 correlates with an influx of new members joining the project and the official 'kick-off meeting'. At this point there is a surge in email communication and a dramatic increase in the number of topics being discussed. Engineering work, much like other problem-solving work (e.g. brainstorming) is often characterised by divergence then convergence of ideas, most notably during the early phases, but even during the late stages of a project. The sudden increase in number of topics at month 5 appears to capture this initial divergence and corresponds to our interviewees' accounts of project activities at this stage. Detection of such patterns may be useful for validating that certain management interventions or actions (such as arranging a kick-off meeting) have generated the desired outcome.

In addition, a normative expectation of project management is that the processes/procedures, including team formation and modes of working, will be established in the early phases, leading to a greater number of management related topics early in the project. For larger projects, such as the one considered in this study, such management activity can occur at each stage as different organisations and sub teams become involved.

The cumulative graph (Fig. 1b) gives a better indication of the growth and expansion in the scope of the project, since the number of topics only increases when vertices (keywords) being added to the graph form new, distinct clusters, rather than joining already existing clusters. Dramatic increases in the cumulative graph trace can therefore either be attributed to topics splitting into sub-topics, or entirely new topics appearing.

Signatures of Scope Creep and Changes in Complexity

Using the size of the topic space as a measure of complexity may be useful as a tool for project managers to monitor and control the work more closely. Rapid growth in topic multiplicity could be an indicator of progress, but also of rapid scope expansion. Most projects face the threat of scope creep, in which new requirements are continually added during development. Managing scope creep is a significant challenge and visualisations such as Figure 1b, which indicate the rate of scope/topic expansion, could provide additional awareness of cases where creep is potentially occurring. Within the engineering project we studied we found documented evidence of the scope being revised, or requests for the original scoping documents to be reviewed (because of concerns about deviations from the original scope) at 8, 10, and 30 months, all of which correspond to periods of growth within Figure 1b. Several interviewed project members also reported that 'scope creep' was a significant issue within this project.

The testing stage of the project signifies an approach towards its overall completion, and a period by which a 'scope freeze' should have already taken place (i.e. a point after which few new topics should emerge). The plateau in the number of topics during this stage (shown in Fig. 1b) seems to indicate that this was the case in this project. However, a continued growth in topics during this phase could indicate complications or increasing complexity.

Signatures of Integrative Work

Since keywords are never removed from the cumulative network, a reduction in the number of clusters in the cumulative graph (Figure 1b) indicates a change in the density of multiple connected clusters, such that the clustering algorithm merges these topics. Two notable occurrences of this topic cluster mergence can be seen at month 18 (labelled as 'Mergence A' on Fig. 1b) and month 33 (labelled as 'Mergence B' on Fig. 1b). From a project management perspective such events are likely to represent substantial synergistic actions within the group. For example, integrating discrete modules of work and making connections between previously disparate topics. For these two decreases in topic cluster counts we were able to find evidence of synergies that instigated such fluctuations. In month 18, an early point in the sub-system testing phase of the project, our analysis led us to discover a number of emails relating to a discussion about employing an analysis technique, 'Failure Mode and Effect Analysis' (FMEA), to a broad spectrum of components that were being engineered as part of the project. Enquiries about the FMEA model had already taken place in the early stages of the project, but without reference to the

components to which it might later be applied. At the point in time that these discussions took place (t=18) the FMEA topic cluster began to merge with topic clusters relating to the parts to which FMEA would be applied, thus reflecting the emerging interaction between the previously disparate aspects of the project.

The other notable topic mergence at month 33 corresponds with the transition from the assembly to the testing phase of the project, where faults were frequently reported, often relating to the interfaces or interactions between separate sub-systems. At this point these sub-systems begin to be viewed as a single system and hence they were frequently referenced together. Being able to spot indicators of synergistic behaviour could provide a useful tool for verifying that the plan is being followed, while a lack of synergy-resembling traces at the expected times could stimulate project members to investigate their absence accordingly.

In the next section we describe our approach for providing indicators of the importance, stability, and maturity of topics that may reveal a potential issue(s), using centrality and density-based strategic diagrams.

6.2 Strategic diagrams for topic clusters

A strategic diagram (shown in Fig. 2) is a two-dimensional planar graph, which places a topic cluster into one of four quadrants according to network-related characteristics, namely its centrality (represented on the x-axis) and density (represented on the y-axis). Strategic diagrams are frequently employed within co-word analyses (e.g. Liu et al., 2014).

A topic cluster's centrality is a measure of the strength of connections to other topic clusters (Callon et al., 1991). While topic clusters are identifiable as discrete entities, their connectedness to other topics can vary. Some are completely isolated, others have connections to many other topic clusters and acquire a position within the centre of the network graph. A topic cluster's density is a measure of the strength of internal connections within the cluster (Callon et al., 1991).

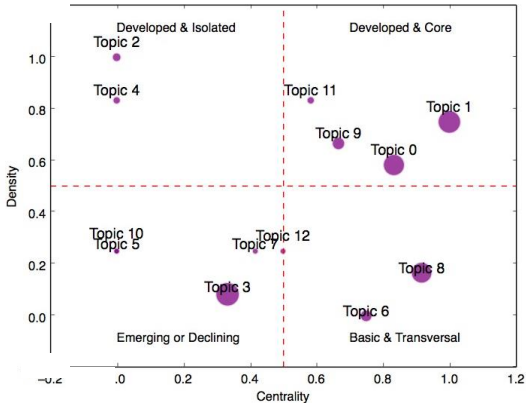

Figure 2. Topics from t=1 plotted on the strategic diagram (according to centrality and density rank)

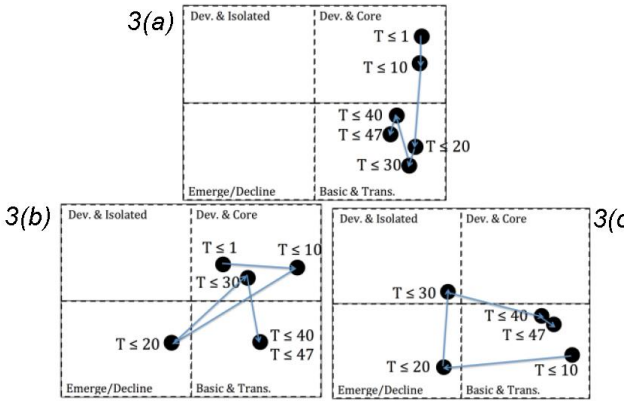

Figure 3. Temporal paths of topics (a) 'Propulsion' topic (b) 'Contract' topic (c) Supplier topic

Topics are plotted on the strategic diagram according to their respective ranks for centrality and density scores. The x and y axis positions represent the median rank (hence topics are evenly distributed either side of each axis). Topic clusters in the upper-right quadrant hold high centrality and high density. These topics are well developed (i.e. words defining the topic are well established and consistently co-occurring) and important for the structuring of the project, since they link to many other topics. Within co-word analysis literature, topics appearing within this quadrant are frequently referred to as core/central topics, or motor-themes (e.g. Coulter et al., 1998, Liu et al., 2014). Topics in the lower-right quadrant are also integral to the project. Their centrality means that they relate to many aspects of the project, however, these topics are not well developed. These are frequently labelled as unstructured, general and transversal topics (e.g. Liu et al., 2014). Topics within the upper-left quadrant have well developed internal ties, but few external ties to other topics and so are less important to the project as a whole. These themes are very specialized and peripheral to the central topics within the project. Themes in the lower-left quadrant are both weakly developed and marginal. The themes of this quadrant have low density and low centrality, mainly representing either emerging or disappearing themes.

Figure 2 shows a snapshot of the strategic diagram representation of the network at t=1 (i.e. month 1 only). The size of each topic bubble within these strategic diagrams corresponds to the frequency with

which words from the topic have been used in the project as a whole. For topic clusters revealed within the first month of the project, we examined their positions within the strategic diagram, determined by their ranked centrality and density scores. We found that Topic 2 (relating to Work Breakdown and Resource Codes) and Topic 4 (relating to the Customer Drawing Register) were well-developed topics, but isolated from the rest of the topics that month. These topics relate to administrative or management activities associated with the project, rather than the central work of the engineers, and as such their position on the periphery of the topic network is to be expected. The work breakdown, resource code and customer drawing register documents are all generated at the beginning of the project, some of which are likely only to be accessed and discussed by administrative or management staff, and all of which are unlikely to be revisited or connected with many other aspects of the project until the work has progressed substantially.

Topic 0 (Outstation Control Console), Topic 1 (Propulsion and Thruster Systems), Topic 9 (Contract Standard Specification Proforma), and Topic 11 (Statutory and Regulatory Compliance Plan) were the developed and core topics within the first month of the project. Much of the work conducted at the beginning of the project was either associated with producing, reviewing and agreeing on a contract and code of practice for the work, or discussing product related topics which were at the core of the project.

Signatures of Coherence and Maturity

The strategic diagram also separates undeveloped, emerging or declining topic clusters on the periphery of the network from those that have stronger connections to other topics (centrality) and greater internal consistency (density). For example Topic 3 from month 1, labelled as a miscellaneous topic (see Table 1) appears within this 'Emerging or Declining' quadrant, and does not clearly represent a single coherent topic. This could be an indicator of a lack of shared understanding or cohesion surrounding this topic, since group members may be using dissimilar terminology to talk about the same subject. Identification of such topics could enable project group members and project managers to review these areas of the project, to ensure that any inconsistencies do not lead to more significant problems. Conversely, developed topics could represent areas where there is greater shared understanding or where the work is more mature.

One might expect a typical path of progression in which new topic emergence is most evident in the early stages of the project, and reduced in the later stages, once the topic space has been defined and work is constrained within this space. The sudden emergence of a new topic at the late stage of a project may therefore be worth highlighting as a potential deviation from this expected progression. Awareness of topics that have not previously been encountered could prompt group members to evaluate the expertise and resources within the group and assess whether they have the capacity to deal with the newly emerging areas of the project. For large organizations such new topics could be used to direct knowledge management efforts, such as the capture of 'Lessons Learned'.

6.3 Changes to the topic structure over time

Using the previously discussed measure of similarity between topic clusters we were able to track clusters with high similarity scores over time, and identify those that made a transition between quadrants of the strategic diagram. Figures 3a, 3b & 3c show the movement of three topic clusters from the cumulative network at intervals $t \leq 1$ (project start), 10, 20, 30, 40 and 47 (project end).

Signatures of Variation in Work Importance

Within this project we would expect all topics related to subsystems (i.e. propulsion, thruster, power, automation systems etc.) to become developed and core, or basic and transversal at some stage during the project, since they are central to the work and relate to many other subsystems. The absence of important sub-systems from these central quadrants could be an indicator of potential issues. We might also expect core topics to be associated with the 'novel' aspects of a project. For many organizations large portions of a project may be repetition of work that they have done before, hence the focus is likely to be on the aspects of the project that present the most novel and challenging work.

We found that our results were in line with these expectations, for example the sub-system relating to 'propulsion' remained a central topic throughout the entire project, only fluctuating in density (reflecting changes in maturity and consistency) over time – as shown in Fig. 3a.

Where management related topics become central to the work, secondary evidence from project documentation and interviews with members of the project revealed that there was a need for

management impetus to control and coordinate the work. This technique therefore provides the potential for automatic indication of the need for management control without direct investigation by the manager. For example, a topic relating to the 'services contract' became core at month 30 (shown in Fig. 3b). All of the interviewed project members reported that unexpected contract issues arose at this time, and reviews of emails sent by the commercial manager revealed that these issues significantly delayed project progress. A corresponding decrease in the number of topics discussed that month can be seen in Fig. 1b (marked as 'contract issues'), indicating a narrower focus within the work. Fig. 3a also shows that contract related topics were core at the early stages of the project (t=1,t=10) however, the importance of forming a contract at this time meant that its presence within this quadrant was likely to be expected.

Signatures of Work Completion or Inattention

We also found evidence of topic clusters moving out of the developed and central quadrants of the diagram and into the emerging/declining quadrant. This transition repeatedly corresponded with interviewees' reported completion of work and the shift away from topics relating to work that had been completed. This is demonstrated in Fig. 3b where the 'contract' topic shifts into the declining themes quadrant (t=20). Once contract related work was completed it was no longer a core focus and did not feature in email discussions. Comparable transitions between quadrants for other management related topics could act as indicators of similar issues. Awareness of topics in decline could be useful for project members to check that discussion has moved on from work that should have been completed, whilst also providing a potential indicator of important topics that should not yet have been concluded, but which are in fact being neglected.

Much like our previous discussion about the potential to spot integrative work, tracking the movement of topics into the low centrality quadrants could be useful for spotting the opposite behaviour, an increase in divergent work on isolated topics. A topic that is not management-related but which shifts to an isolated position within the network may need to be highlighted in order to check that it is still aligned with other aspects of the project. One such example is shown in Fig. 3c. This topic relates to a particular supplier (anonymised). Towards the end of the assembly phase of the project (t=30) some project members found themselves unable to use their previous contact point for this supplier, and hence discussion focused on the issue of re-establishing contact, rather than being related to specific areas of project work involving the supplier. Topics that remain within an isolated position for a period of time over a threshold might be an indicator of an unresolved issue and could prompt closer investigation from project members. Patterns of movement within the strategic diagrams could form useful templates for successful/unsuccessful projects, against which future projects could be compared.

7 CONCLUSION

To reiterate our main findings, we have uncovered evidence of patterns and trends (which we refer to as *signatures*), relating to the topics revealed by email subject lines, which correspond to significant events or issues within the engineering project studied (see Table 2).

Table 2. Summary of e-mail topic signatures

Signature	Interpretation
Increase in topic count within the single month co-word network	Corresponding increase in work complexity and breadth of focus
Decrease in topic count within the single month co-word network	Consolidation work around the time of a phase transition
Step changes in topic count for single month network	Significant shift in focus / Project phase transition
Increase in topic count within the cumulative co-word network	Growth and expansion in the scope of the project / Divergence in work directions
Increase in cumulative topic count during late stages of the project	Occurrence of scope creep or work complications
Decrease in no. of topics within the cumulative co-word network	Integrative work combining discrete aspects of the project / Group synergy
Topic position within high centrality quadrants of the strategic diagram	Important and core aspects of the project work

Topic position within high density quadrants of the strategic diagram	Coherent, developed and mature aspects of the work
Topic position within low centrality quadrants of the strategic diagram	Isolated and disconnected aspects of the work / Work completion or inattention
Topic position within low density quadrants of the strategic diagram	Lack of cohesion and shared understanding

These signatures enable the identification of project transitions, divergence and shifts in focus, scope creep and changes in complexity, integrative work, topic coherence and maturity, and variation in importance. Within the project studied, many signatures corresponded with points at which there was a need for management input to control the work. We argue, therefore, that integration of the co-word analysis technique into project management tools has potential to be useful for increasing project members' awareness of such events, and providing a mechanism to prompt managers to investigate emerging issues. Additionally, some of the signatures corresponded with various stages of progression through a project, such as transitions between project phases, and hence provided information that could be used to validate progress against a schedule or plan. Furthermore, the highly automatable nature of the analysis increases its usefulness, by reducing the costs associated with manual project reporting and feedback mechanisms.

Since our analysis focuses on a single project, further validation work is required in order to provide evidence that these signatures are generalizable across projects, organizations and different types of collaborative work and to evaluate the utility of the information that this analysis provides in the hands of project members during a 'live' project.

8 REFERENCES

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 2008(10).

Callon, M., Courtial, J. P., & Laville, F. (1991). Co-word analysis as a tool for describing the network of interactions between basic and technological research. Scientometrics, 22(1), 155-205.

Callon, M., Courtial, J. P., Turner, W. A., & Bauin, S. (1983). From translations to problematic networks: An introduction to co-word analysis. Social science information, 22(2), 191-235.

Coulter, N., Monarch, I., & Konda, S. (1998). Software engineering as seen through its research literature. Journal of the American Society for Information Science, 49(13), 1206-1223.

Cselle, G., Albrecht, K., & Wattenhofer, R. (2007). BuzzTrack: topic detection and tracking in email. In Proc. IUI (pp. 190-197).

Dredze, M., Lau, T., & Kushmerick, N. (2006). Automatically classifying emails into activities. In Proc. IUI (pp. 70-77).

Ducheneaut, N., & Bellotti, V. (2001). Email as Habitat: An Exploration of Embedded Personal Information Management. Interactions, 8(5), 30-38.

Gopsill JA, Payne SJ & Hicks BJ (2013) "An Exploratory Study into Automated Real-Time Categorisation of Engineering E-Mail", IEEE International Conference on Systems, Man, and Cybernetics, Manchester, UK.

Gopsill, J., Jones, S., Snider, C., Shi, L., McMahon, C., Hicks, B. J. (2014). Understanding the engineering design process through the evolution of engineering digital objects. DESIGN 2014: 13th International Design Conference.

Grønbæk, K., Kyng, M., & Mogensen, P. (1993). CSCW challenges: cooperative design in engineering projects. CACM, 36(6), 67-77.

Grudin, J. (1988). Why CSCW applications fail: problems in the design and evaluation of organizational interfaces. In Proc. In Proc. CSCW (pp. 85-93).

Hicks, B. (2013). The Language of Collaborative Engineering projects. International Conference on Engineering Design (ICED13), August 19-22nd 2013, Seoul, Korea.

Jones, S., Payne, S., Hicks, B. and Watts, L. (2013). Visualization of Heterogeneous Text Data in Collaborative Engineering Projects. The 3rd IEEE Workshop on Interactive Visual Text Analytics. IEEE VIS 2013.

Korba, L., Song, R., Yee, G., & Patrick, A. (2006). Automated social network analysis for collaborative work.

Liu, L. C., & Horowitz, E. (1989). A formal model for software project management. IEEE Transactions on Software Engineering, 15(10), 1280-1293.

Liu, Y., Goncalves, J., Ferreira, D., Xiao, B., Hosio, S., & Kostakos, V. (2014). CHI 1994-2013: Mapping two decades of intellectual progress through co-word analysis. In Proc. CHI.

Loftus, C., McMahon, C. and Hicks, B., (2008). Issues and challenges for improving email use in engineering design. In: NordDesign 2008, 2008-08-21 - 2008-08-23, Tallinn University of Technology, Tallinn.

Loftus, C., Hicks, B. J., & McMahon, C. A. (2010). Understanding the use of email in engineering: A scenario based approach. In 11th International Design Conference, DESIGN 2010. (pp. 1575-1584)

Mackay, W. E. (1988). More than just a communication system: diversity in the use of electronic mail. In Proc. CSCW (pp. 344-353).
Porter, M. (2009). {The Porter Stemming Algorithm}.
Shi L., Gopsill J., Snider C., Jones, S., Newnes, L. and Culley, S. (2014). Towards identifying patterns in engineering documents to aid project planning. DESIGN 2014: 13th International Design Conference. Dubrovnik, Croatia.
Snider, C. M., Jones, S., Gopsill, J., Shi, L., Hicks, B. J. (2014). A framework for the development of characteristic signatures of engineering projects. DESIGN 2014: 13th International Design Conference. Dubrovnik, Croatia.
Surendran, A. C., Platt, J. C., & Renshaw, E. (2005). Automatic Discovery of Personal Topics to Organize Email. In CEAS.
Wasiak, J., Hicks, B., Newnes, L., Dong, A., & Burrow, L. (2010). Understanding engineering email: the development of a taxonomy for identifying and classifying engineering work. Research in Engineering Design, 21(1), 43-64.
Wasiak, J. O., (2010). A Content Based Approach for Investigating the Role and Use of Email in Engineering Design Projects.
Wolf, T., Schroter, A., Damian, D., & Nguyen, T. (2009). Predicting build failures using social network analysis on developer communication. In Proc. 31st International Conference on Software Engineering (pp. 1-11).

9 ACKNOWLEDGMENTS

The research reported in this paper is funded by Engineering and Physical Sciences Research Council (EP/K014196/1).
Due to license restrictions, supporting data are not openly available